FAIR as a factor for bigger research impact

Guiding principles for successful data sharing and reuse

Wolmar Nyberg Åkerström
ELIXIR Sweden – NBIS / SciLifeLab
data@nbis.se

Why talk about data sharing?

- Policymakers are **pushing for research data to be made available** as openly as possible
- Big investments are being made in **infrastructure and skills for data sharing and reuse**
- Some motivating factors
  - Democratic principles
  - Good research practices
  - Societal and academic impact

Swedish Research Bill 2021–2024*

> *[…] research data shall be made accessible as* ***open as possible and as closed as necessary***

* Our translation from Swedish

The EU's Open Science policy

> ***FAIR** […]* ***open data sharing should become the default*** *for the results of EU-funded scientific research*

What’s in it for the individual?

- Successfully sharing data can **increase the citation rate** of your articles
- By striving for FAIR you **make it easier for your future self and collaborators** to use the data your are producing now
- Be **prepared for future opportunities** with stricter data sharing requirements

> “*[…] papers with publicly available microarray data* ***received more citations*** *than similar papers that did not make their data available, even after controlling for many variables known to influence citation rate. We found the* ***open data citation benefit*** *for this sample to be 9% overall*
>
> – Piwowar, H. A., & Vision, T. J. (2013)

Piwowar, H. A., & Vision, T. J. (2013). Data reuse and the open data citation advantage. *PeerJ*, *1*, e175. doi:10.7717/peerj.175

Reuse by people and machines

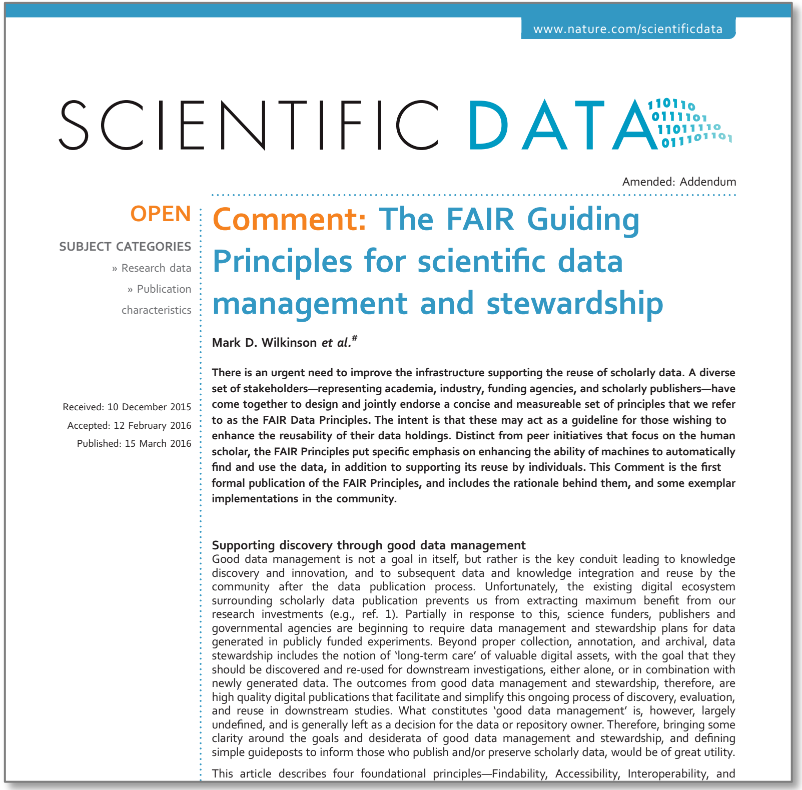
More value to publicly funded research
Improve peer-review process
Works with my software
High research integrity
Better research output
M
FAIR

Ronot: MetaManMachine by Nikola Vasiljevic (2021), CC BY-SA 4.0, doi:10.5281/zenodo.4471098

Background from ‘FAIR Principles’ by Martínez-Lavanchy, et al (2019), CC-BY 4.0. doi:10.11581/dtu:00000049

The FAIR principles

- Promote **efficient data discovery and reuse** by providing guidelines to make digital resources
  - ☐ Findable
  - ☐ Accessible
  - ☐ Interoperable
  - ☐ Reusable
- Address aspects **enabling software and infrastructure** to automatically find and use research data



www.nature.com/scientificdata

SCIENTIFIC DATA

Amended: Addendum

OPEN

SUBJECT CATEGORIES
» Research data
» Publication characteristics

Comment: The FAIR Guiding Principles for scientific data management and stewardship

Mark D. Wilkinson *et al.*#

There is an urgent need to improve the infrastructure supporting the reuse of scholarly data. A diverse set of stakeholders—representing academia, industry, funding agencies, and scholarly publishers—have come together to design and jointly endorse a concise and measureable set of principles that we refer to as the FAIR Data Principles. The intent is that these may act as a guideline for those wishing to enhance the reusability of their data holdings. Distinct from peer initiatives that focus on the human scholar, the FAIR Principles put specific emphasis on enhancing the ability of machines to automatically find and use the data, in addition to supporting its reuse by individuals. This Comment is the first formal publication of the FAIR Principles, and includes the rationale behind them, and some exemplar implementations in the community.

Received: 10 December 2015
Accepted: 12 February 2016
Published: 15 March 2016

Supporting discovery through good data management

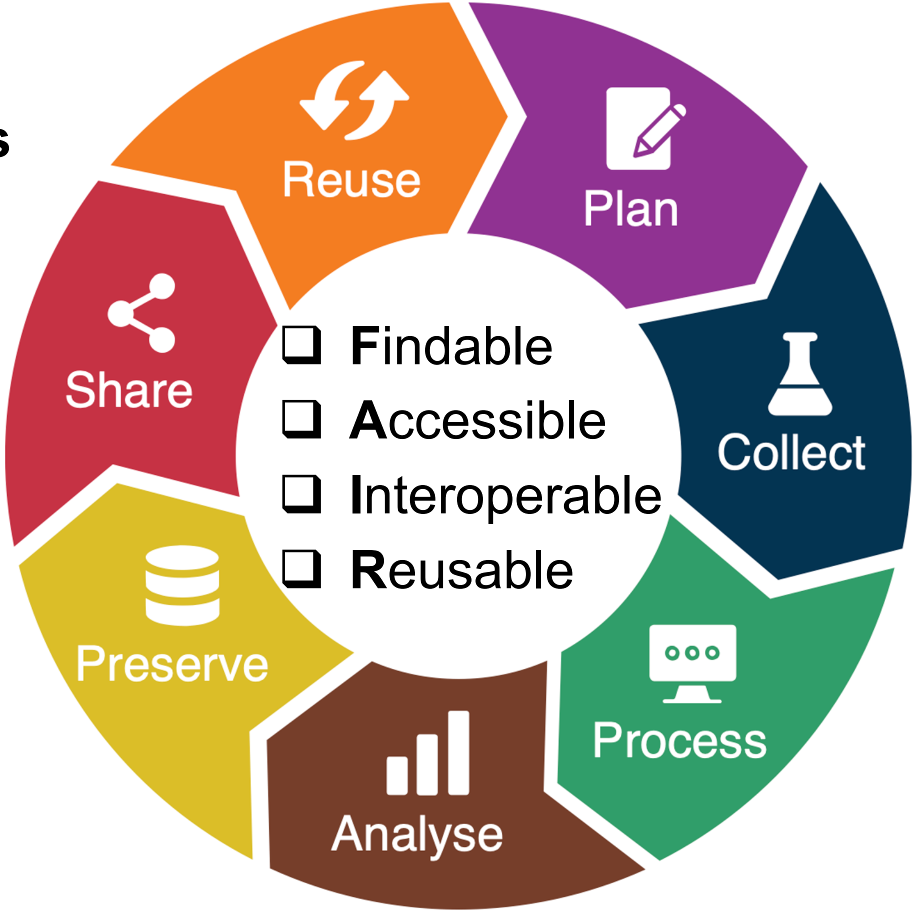
Good data management is not a goal in itself, but rather is the key conduit leading to knowledge discovery and innovation, and to subsequent data and knowledge integration and reuse by the community after the data publication process. Unfortunately, the existing digital ecosystem surrounding scholarly data publication prevents us from extracting maximum benefit from our research investments (e.g., ref. 1). Partially in response to this, science funders, publishers and governmental agencies are beginning to require data management and stewardship plans for data generated in publicly funded experiments. Beyond proper collection, annotation, and archival, data stewardship includes the notion of 'long-term care' of valuable digital assets, with the goal that they should be discovered and re-used for downstream investigations, either alone, or in combination with newly generated data. The outcomes from good data management and stewardship, therefore, are high quality digital publications that facilitate and simplify this ongoing process of discovery, evaluation, and reuse in downstream studies. What constitutes 'good data management' is, however, largely undefined, and is generally left as a decision for the data or repository owner. Therefore, bringing some clarity around the goals and desiderata of good data management and stewardship, and defining simple guideposts to inform those who publish and/or preserve scholarly data, would be of great utility.

This article describes four foundational principles—Findability, Accessibility, Interoperability, and

Wilkinson et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*(1), 160018. doi:10.1038/sdata.2016.18

A FAIR data lifecycle

- The FAIR principles relies on **good data management practices** in all phases of research
  - Ethics and legislation
  - Information security
  - Research documentation
  - Data organisation
- **FAIR data ≠ Open data**
  Carefully consider what data and versions to preserve and under what conditions they will be shared



☐ **F**indable
☐ **A**ccessible
☐ **I**nteroperable
☐ **R**eusable

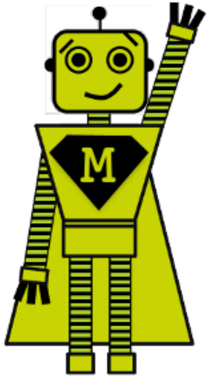
Findable

The first step in (re)using data is to find them. It should be easy for both humans and computers.

- ☐ You can **identify the data** and rely on that identification to find the data in the far future
- ☐ You can find the data when you **search for them by their descriptive attributes**

Guiding principles

- Make an effort to describe the data with **rich metadata**
- Assign a **unique and persistent identifier (PID)**, such as a DOI, and include it in the metadata
- Ensure that data are **findable using a search service**

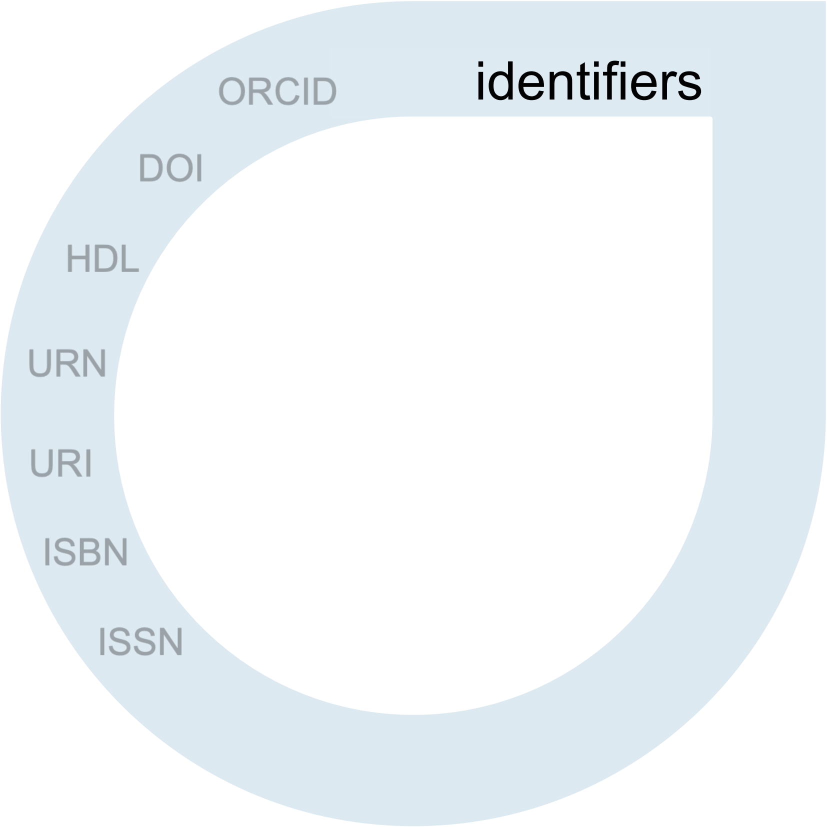


MetaManMachine by Nikola Vasiljevic (2021), CC BY-SA 4.0, doi:10.5281/zenodo.4471098

Unique & persistent identifiers

Identifiers are labels that help us **uniquely identify physical and digital objects** and services

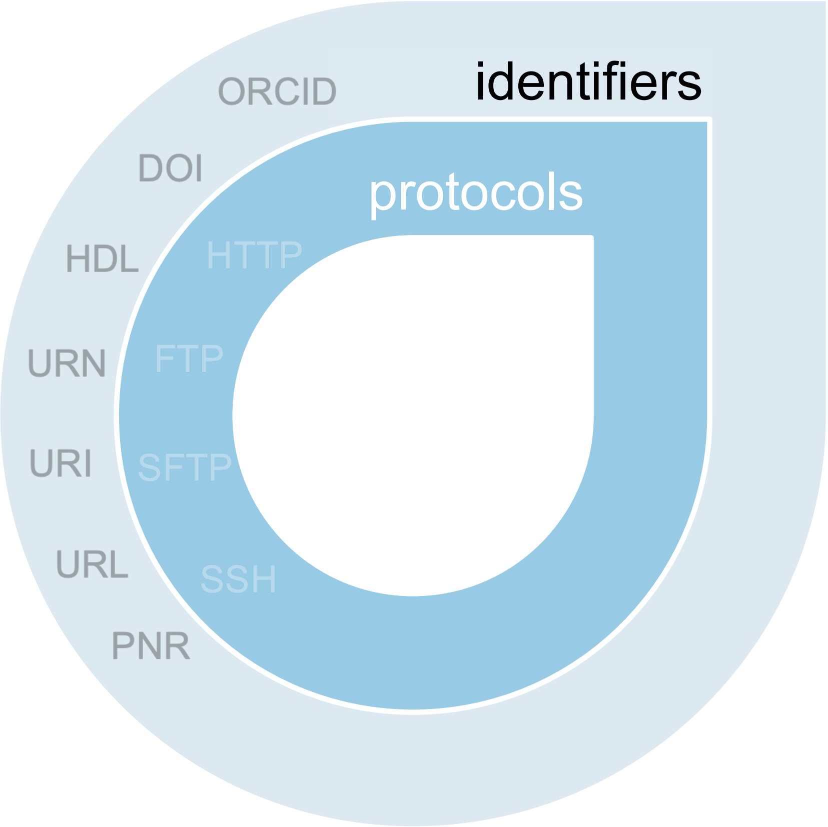
- We ideally want an identifier to be globally unique, persistent (never a broken link), and resolvable
- A DOI meets all the requirements, so does URN, ORCID, etc.
- **DOI:** 10.5281/zenodo.4471098
  Maintained by Zenodo
  https://doi.org/10.5281/zenodo.4471098



Access protocols

Protocols are **procedures to access** data and metadata

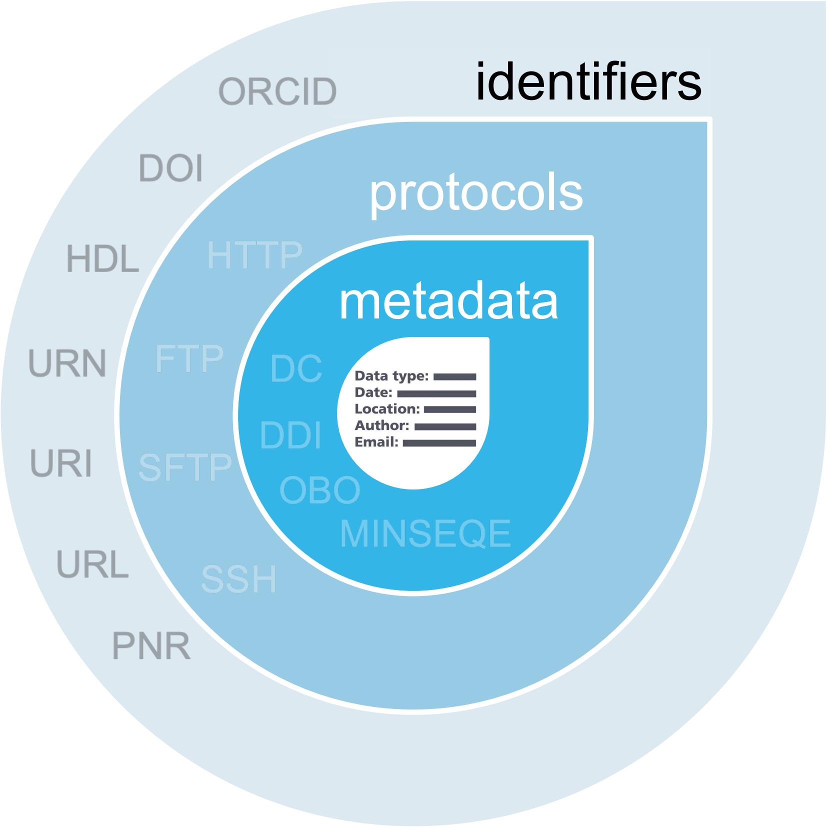
- Prefer well-established protocols, such as http(s) and ftp. This is how the web browser connect to the Internet and downloads a web page.
- Fully automated access to sensitive data might not be possible. Provide contact details and clear instructions.



Metadata

Metadata is data that defines
and **describes other data**
or resources

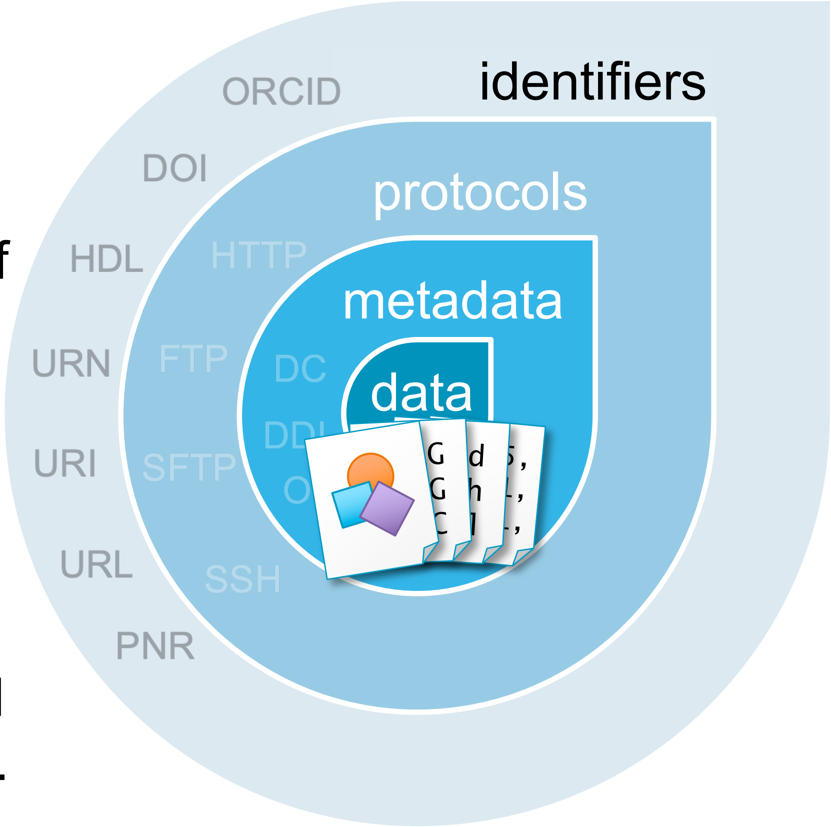
- Metadata is everywhere. From a file name and the file extension to embedded camera details in a picture to a data table listing tissue samples and their properties.
- Use widely-adopted templates and vocabularies so others understand their structure and meaning



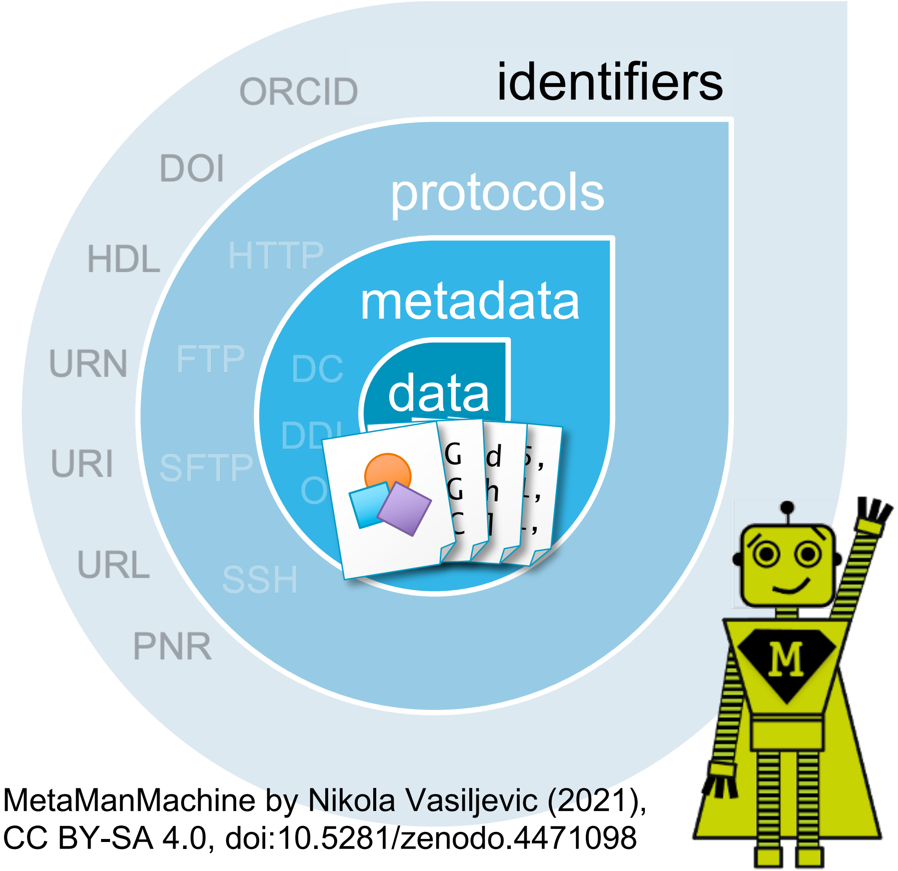
Data and data files

Data (and metadata) **represent observations,** measurements and information that support your research

- All digital file formats are in danger of becoming outdated. If that happens, future software may not be able to read or show the content correctly.
- You should choose a file format that is likely to be usable in the future.
- Document the content structures and relationships between your data files.



Data as a digital resource



MetaManMachine by Nikola Vasiljevic (2021),
CC BY-SA 4.0, doi:10.5281/zenodo.4471098

Accessible

Once identified, we need to know how to locate and access them.

- ☐ You can **obtain the data** from wherever it is stored
- ☐ You can issue a **request to get access** if the data cannot be shared openly

www.go-fair.org

@GOFAIRofficial

Guiding principles

- Choose an identifier that can be used to **access the data using a widely adopted communications protocol**
- Provide **a means to request the data** if access is restricted
- **Keep the metadata available**, even if the data is removed

Interoperable

Working with someone else files can be challenging. We need to find out how to integrate the data in our own workflows

☐ You can **open the data files, read their content and understand** what the data represent

Guiding principles

- Use a formal, accessible, shared, and broadly applicable language for knowledge representation
- Use vocabularies that follow FAIR principles
- Include PIDs and links to required resources

Reusable

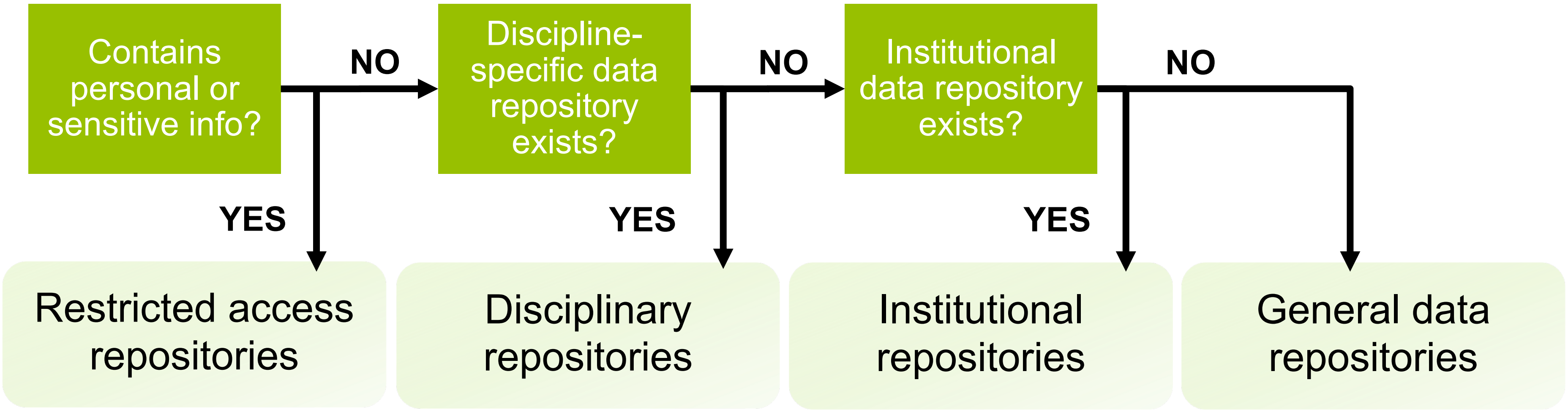
To optimise for reuse, metadata and data should be well-described and accessible to machines

- ☐ You can determine under what conditions and **to what extent the data can be used** in your project
- ☐ You can **combine the data with other sources** with acceptable effort
- ☐ You can **cite and reference** the data

Guiding principles

- Describe the context, structure and content of the data, linking documentation, protocols and papers to the data
- Use well-established and sustainable file formats and descriptive standards
- Clearly state under which license the data can be used

Selecting a data repository



Contains personal or sensitive info?

NO

Discipline-specific data repository exists?

NO

Institutional data repository exists?

NO

YES

YES

YES

Restricted access repositories

Disciplinary repositories

Institutional repositories

General data repositories

EUROPEAN GENOME-PHENOME ARCHIVE

ENA European Nucleotide Archive

PRIDE

BioSamples

Protein Data Bank in Europe
Bringing Structure to Biology

Stockholm University

SND Svensk nationell datatjänst

SciLifeLab Data Repository

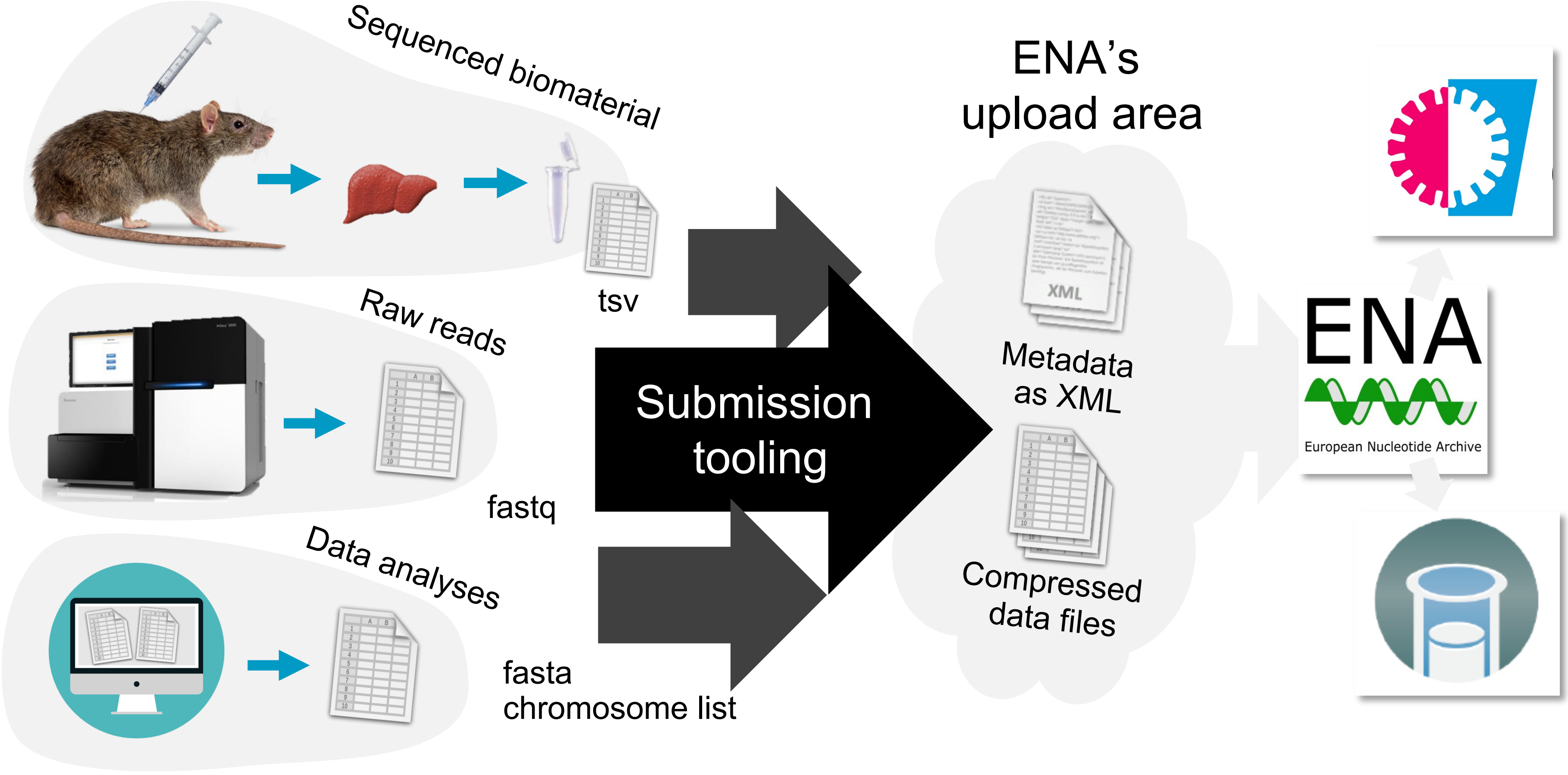
zenodo

figshare

EUDAT

https://www.re3data.org/ https://scilifelab-data-guidelines.readthedocs.io/

Discipline specific (ENA)



Sequenced biomaterial

tsv

Raw reads

fastq

Data analyses

fasta
chromosome list

Submission
tooling

ENA's
upload area

Metadata
as XML

Compressed
data files

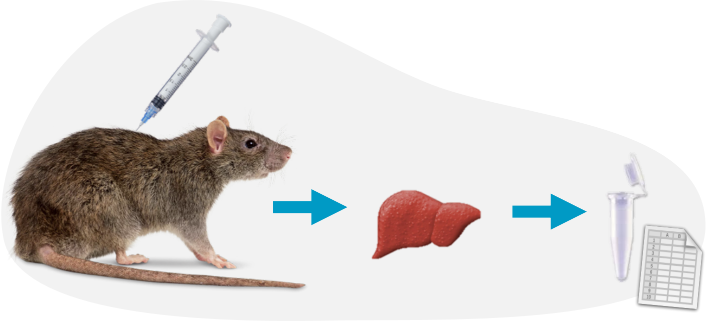
ENA
European Nucleotide Archive

Sample: Sequenced Biomaterial

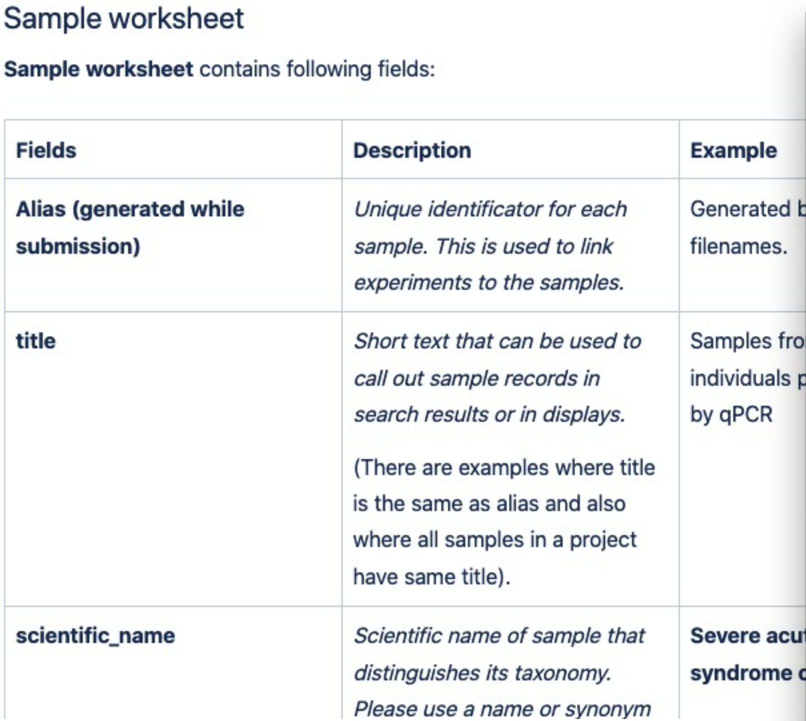
- Should have a unique name that can identify the source material
- Always associated with a taxon with further descriptions using fields from a curated checklist
- ENA virus pathogen reporting standard checklist has 35 fields
  - 9 required
  - 15 recommended
  - 11 optional

Info should be available early on

- A sample name
- Associated taxon
- A sample description
- Use the checklist: **ERC000033**



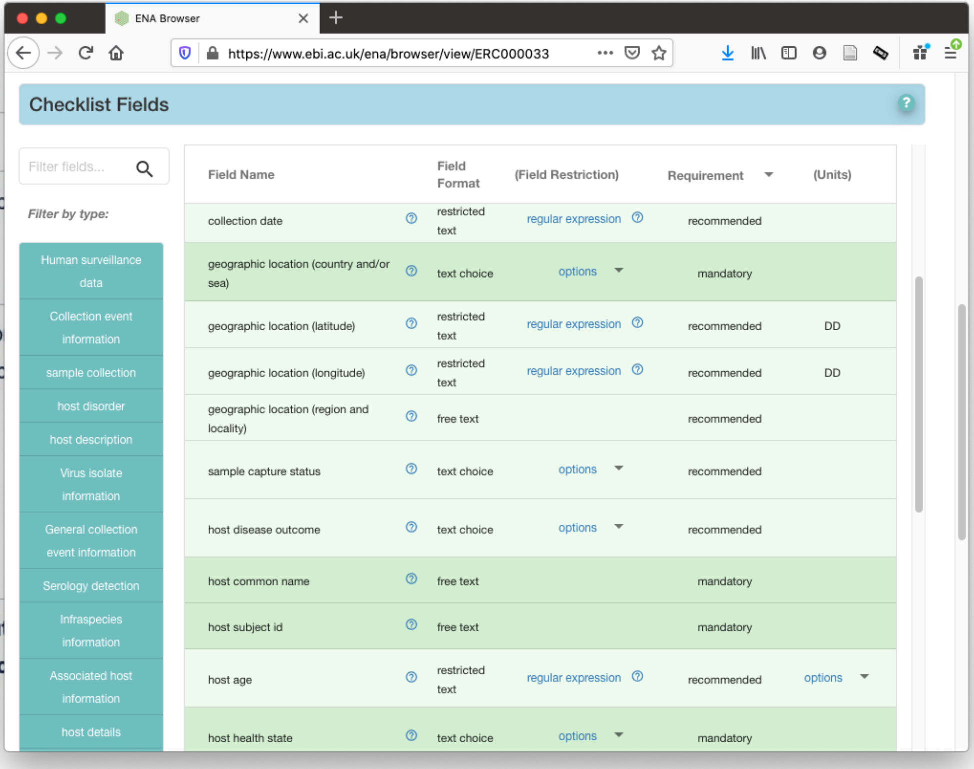
Checklist: ERC000033



Sample worksheet

Sample worksheet contains following fields:

Fields	Description	Example
Alias (generated while submission)	*Unique identificator for each sample. This is used to link experiments to the samples.*	Generated b filenames.
title	*Short text that can be used to call out sample records in search results or in displays.* (There are examples where title is the same as alias and also where all samples in a project have same title).	Samples fro individuals p by qPCR
scientific_name	*Scientific name of sample that distinguishes its taxonomy. Please use a name or synonym*	**Severe acu** **syndrome c**



Checklist Fields

Field Name	Field Format	(Field Restriction)	Requirement	(Units)
collection date	restricted text	regular expression	recommended	
geographic location (country and/or sea)	text choice	options	mandatory	
geographic location (latitude)	restricted text	regular expression	recommended	DD
geographic location (longitude)	restricted text	regular expression	recommended	DD
geographic location (region and locality)	free text		recommended	
sample capture status	text choice	options	recommended	
host disease outcome	text choice	options	recommended	
host common name	free text		mandatory	
host subject id	free text		mandatory	
host age	restricted text	regular expression	recommended	options
host health state	text choice	options	mandatory	

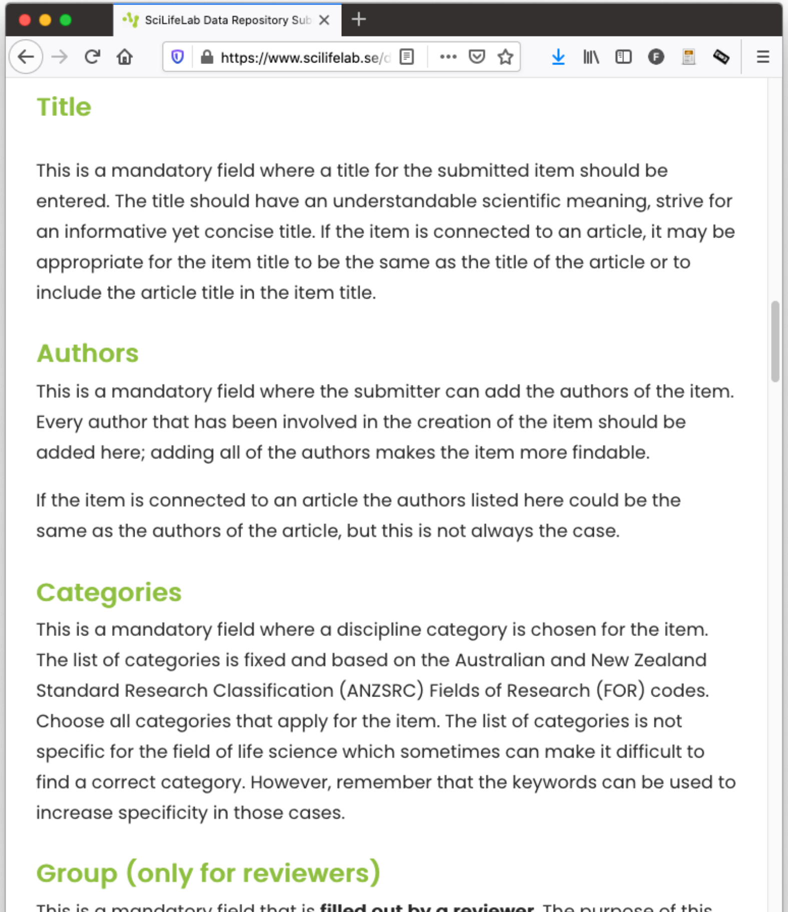
https://www.ebi.ac.uk/ena/browser/view/ERC000033

dx.doi.org/10.17504/protocols.io.bh5dj826

Institutional (SciLifeLab)

Consult the SciLifeLab Data Repository **Submission Guidelines**

- ❑ Choose file formats that are likely to be **usable in the future**
- ❑ Use **widely-adopted templates and vocabularies** where possible
- ❑ Add a **Readme to store a copy of the metadata** and from the Figshare
- ❑ Add a **Manifest** for users to verify that nothing has been lost in transit



Title

This is a mandatory field where a title for the submitted item should be entered. The title should have an understandable scientific meaning, strive for an informative yet concise title. If the item is connected to an article, it may be appropriate for the item title to be the same as the title of the article or to include the article title in the item title.

Authors

This is a mandatory field where the submitter can add the authors of the item. Every author that has been involved in the creation of the item should be added here; adding all of the authors makes the item more findable.

If the item is connected to an article the authors listed here could be the same as the authors of the article, but this is not always the case.

Categories

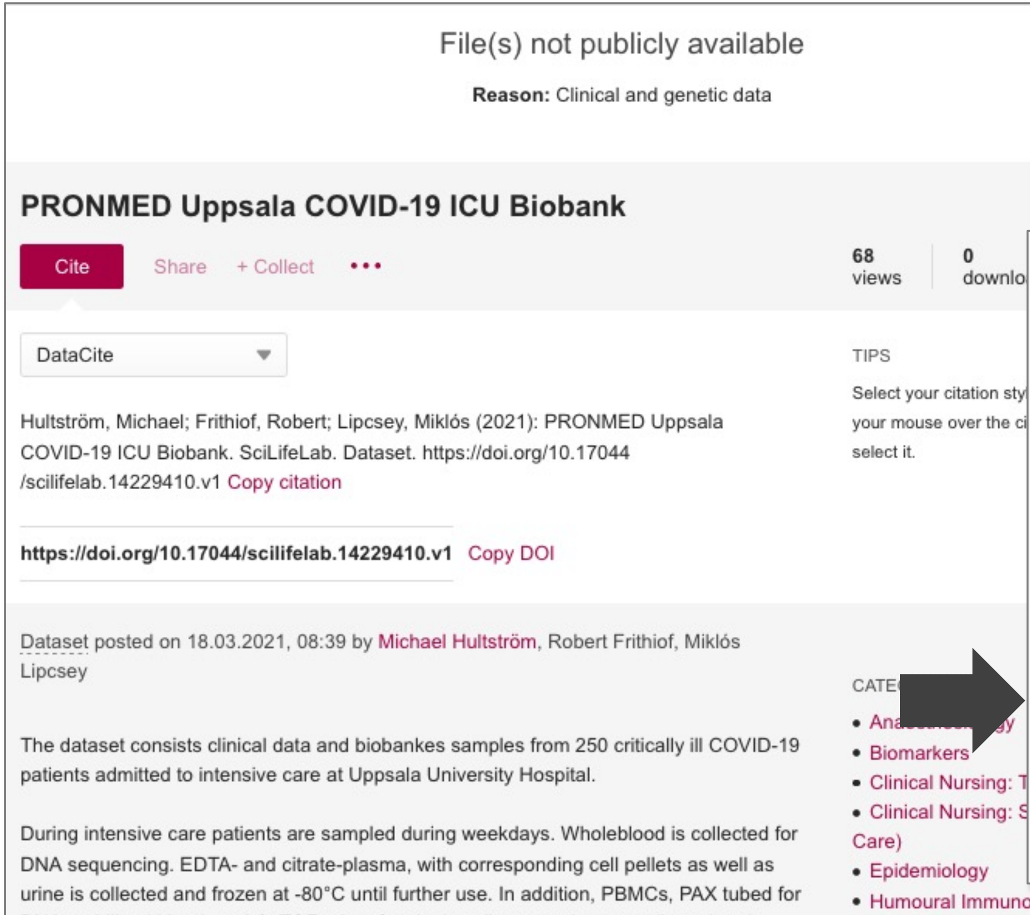
This is a mandatory field where a discipline category is chosen for the item. The list of categories is fixed and based on the Australian and New Zealand Standard Research Classification (ANZSRC) Fields of Research (FOR) codes. Choose all categories that apply for the item. The list of categories is not specific for the field of life science which sometimes can make it difficult to find a correct category. However, remember that the keywords can be used to increase specificity in those cases.

Group (only for reviewers)

This is a mandatory field that is **filled out by a reviewer**. The purpose of this

Restricted access?

Cite, reference
and let others know
that the data exists



File(s) not publicly available

Reason: Clinical and genetic data

PRONMED Uppsala COVID-19 ICU Biobank

TIPS

Select your citation sty
your mouse over the ci
select it.

Hultström, Michael; Frithiof, Robert; Lipcsey, Miklós (2021): PRONMED Uppsala COVID-19 ICU Biobank. SciLifeLab. Dataset. https://doi.org/10.17044/scilifelab.14229410.v1 Copy citation

https://doi.org/10.17044/scilifelab.14229410.v1 Copy DOI

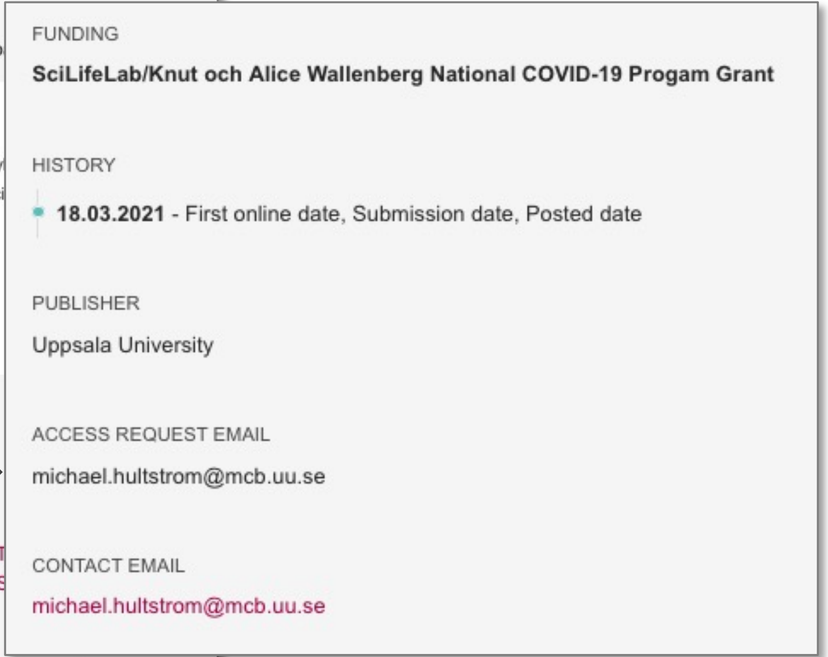
Dataset posted on 18.03.2021, 08:39 by Michael Hultström, Robert Frithiof, Miklós Lipcsey

The dataset consists clinical data and biobankes samples from 250 critically ill COVID-19 patients admitted to intensive care at Uppsala University Hospital.

During intensive care patients are sampled during weekdays. Wholeblood is collected for DNA sequencing. EDTA- and citrate-plasma, with corresponding cell pellets as well as urine is collected and frozen at -80°C until further use. In addition, PBMCs, PAX tubed for

CATE

- Ana gy
- Biomarkers
- Clinical Nursing: T
- Clinical Nursing: S Care)
- Epidemiology
- Humoural Immunology and



FUNDING

SciLifeLab/Knut och Alice Wallenberg National COVID-19 Progam Grant

HISTORY

- **18.03.2021** - First online date, Submission date, Posted date

PUBLISHER

Uppsala University

ACCESS REQUEST EMAIL

michael.hultstrom@mcb.uu.se

CONTACT EMAIL

michael.hultstrom@mcb.uu.se

Good practices

☐ **Secure/organise data & analyses**, by managing back-ups, access restrictions, versioning, docs, scripts and transcripts

☐ **Deposit and share data** using restricted or public access data repositories that promote FAIR data principles

☐ **Adhere to community standards**, such as file formats, data dictionaries, controlled vocabularies and metadata

☐ **Maintain a Data Management Plan,** outlining the project's data management practices

Contact us

- nbis.se
- scilifelab.se/data
- data@nbis.se
- wolmar.n.akerstrom@nbis.se